JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 19, NO. 4 WINTER 1982

A LATENT TRAIT METHOD FOR DETERMINING INTRAJUDGE INCONSISTENCY IN THE ANGOFF AND NEDELSKY TECHNIQUES OF STANDARD SETTING

WIM J. VAN DER LINDEN
Twente University of Technology
Enschede, The Netherlands

In objectives-based instructional programs, absolute rather than relative standards are used to evaluate student performances. These standards can be considered to be the translation of the learning objectives in the program into cutoff scores on the true score scale of the test. They constitute the predetermined levels that the student's true performance must exceed to be granted mastery status and, for instance, to be allowed to proceed with the next instructional unit. An important problem in mastery testing is how to set standards separating those students who meet learning objectives from those who do not. Several techniques of standard setting have been proposed, and an extensive literature on the problem is available which has recently been reviewed by several authors (Glass, 1978; Hambleton, 1980; Hambleton, Powell & Eignor, 1979; Jaeger, 1979; Shepard, 1980a, 1980b). In this paper, the emphasis will be on the Angoff (1971) and Nedelsky (1954) techniques; these are commonly classified as techniques based on judgement of test content. A review of the mastery testing literature is given in Hambleton, Swaminathan, Algina, and Coulson (1978) and van der Linden (1982).

It has been argued that all standard setting is arbitrary (Glass, 1978; Shepard, 1979, 1980a, 1980b). This is correct since standards ought to reflect learning objectives and these ultimately rest on value judgments and norms. In addition, the various standard setting techniques available differ in varying degree according to the conception of mastery underlying the way standards are obtained. Therefore, different results can be expected both for different techniques and for different persons using the same technique. This has been confirmed in many experiments (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Koffler, 1980; Saunders, Ryan & Huynh, 1981; Skakun & Kling, 1980). That all standard setting is ultimately arbitrary led Glass to the pessimistic conclusion that we should abandon the use of these techniques. However, as Hambleton (1978) and Popham (1978) have put forward, arbitrariness does not necessarily have a negative meaning. There are many other instances in which arbitrary choices have to be made in which deliberate, defensible results are obtained. What should be avoided is capricious standard setting, that is, standard setting in which the learning objectives are inconsistently translated into the cutoff score and, in fact, erratic standards of mastery are obtained.

In a sense, the present paper addresses this second, negatively loaded type of arbitrariness. Its concern is not with differences in standard setting between persons or techniques. There simply is no reason to expect the same result when persons with different interpretations of learning objectives or techniques based on different concep-

The author is indebted to Ronald K. Hambleton and Gideon J. Mellenbergh for their comments. Thanks are also due to Ronny F. A. Wierstra and the PLON-CITO team for participating in the empirical study and to Paula Achterberg for typing the manuscript.

tions of mastery are used. Rather, the interest is in the occurrence of *intrajudge inconsistency* when the Angoff or Nedelsky technique is used to set a standard for a given test. Intrajudge inconsistency arises when the judge specifies probabilities of success on the items which are incompatible with each other and, consequently, imply different standards. An example is an Angoff judge assigning a low probability of success for a borderline student on an easy item but a large probability on a difficult item. Obviously, these two judgments are inconsistent; the former implies a low standard whereas the latter indicates that a high standard should be set. Inconsistencies may be caused by items being perceived differently from the way they actually function. The concept of intrajudge inconsistency will be elucidated further below.

Three possible sources of inconsistency in standard setting were distinguished in the preceding: (1) inconsistency due to different conceptions of mastery underlying the technique, (2) interjudge inconsistency due to different interpretations of learning objectives, and (3) intrajudge inconsistency. So far, no attention has been paid to the last (intrajudge) source of inconsistency, and results of the Angoff or Nedelsky technique are generally employed without checking the consistency of the judge. That intrajudge inconsistency has eluded the attention of researchers may be related to the fact that classical test theory does not provide methods for analyzing probabilities of successful item responses as a function of the mastery level of the student. The idea of item-observed score regression comes closest to what we need. However, the use of this regression function may give rise to misleading results, not only as a consequence of the fact that test scores are unreliable but also because they are based in part on the item response in question. (See, for example, Lord, 1980, sect. 3.1.)

It is the intent of this paper to show how a method for analyzing intrajudge inconsistency can be derived from latent trait theory (Birnbaum, 1968; Lord, 1980; Wright & Stone, 1979) and how a simple index of consistency can be defined which can be used for deciding whether standards have been set consistently. The method can be used, for example, to select judges, to evaluate training programs for judges, or to assess consequences of modifying standard-setting techniques. Before introducing the method, however, the Angoff and Nedelsky standards will be discussed following a slightly different notation so that some of their properties can be indicated and the possibility of latent trait analysis becomes obvious. Finally, the paper presents results from an empirical investigation which demonstrate how the method should be used and shed some light on the question of how consistently the Angoff and Nedelsky techniques are used in a typical educational situation.

THE ANGOFF AND NEDELSKY TECHNIQUES

Although the Angoff technique was introduced only in a short footnote (Angoff, 1971, p. 515), it has become one of the best known and widely used methods of standard setting. It is suited for dichotomously scored items and consists of the following few steps: A content specialist is asked to imagine a student just meeting the requirements as formulated in the learning objectives. This may be a hypothetical as well as a real student. Keeping this borderline student in mind, he/she is requested to inspect the test, item by item, and to specify for each item the probability that the student will answer it correctly. The standard is equal to the sum of the probabilities.

Let $P_i(+|\theta)$ be the probability that a student with mastery level θ answers item i correctly, and let θ_c denote the mastery level of the borderline student whom the content

specialist has in mind. The term mastery level will be used in the paper as a generic term for the degree to which the student masters the domain of knowledge or skill formulated in the learning objectives. It will be assumed that the domain is homogeneous enough to conceive of the mastery level θ as a unidimensional attribute. In the Angoff method, $P_i(+\,|\,\theta_c)$ is specified for each of the n items in the test and the standard is defined by:

$$\sum_{i=1}^{n} P_i(+\,|\,\theta_c). \tag{1}$$

Note that $P_i(+\,|\,\theta)$ is not only the probability of success but also the expected item score for a student with mastery level θ, since it holds that:

$$E(u_i\,|\,\theta) = 1.P_i(+\,|\,\theta) + 0.[1 - P_i(+\,|\,\theta)], \tag{2}$$

where $u_i = 0, 1$ denotes the item score. In classical test theory, the true number-right score for a fixed person τ, is defined as the expected value of his/her observed test score $X = \Sigma_{i=1}^{n} u_i$. From Equations 1 and 2, it follows that:

$$\sum_{i=1}^{n} P_i(+\,|\,\theta_c) = \sum_{i=1}^{n} E(u_i\,|\,\theta_c) = E\left(\sum_{i=1}^{n} u_i\,|\,\theta_c\right) = E(X\,|\,\theta_c) \equiv \tau_c. \tag{3}$$

Thus, the Angoff technique translates the performance of a borderline student who just meets the learning objectives into a cutoff score on the *true* score scale of the given test. This cutoff score can subsequently be used to determine an optimal cutoff score on the observed score variable, preferably using decision-theoretic procedures (van der Linden, 1980). For future reference, it is noted that the relation given in Equation 3 is known as the test characteristic curve (Lord, 1980, p. 49).

The Nedelsky technique was introduced some 25 years before the Angoff technique became known (Nedelsky, 1954). It is also based on judgment of test content and uses the same setting of judges who, imagining a borderline student, are requested to go through the test item by item. However, it can only be used for multiple-choice items and is based on an all-or-none model with respect to the item alternatives. It assumes that a student knows which alternatives are incorrect and guesses between the remaining alternatives (if more than one are left). It is the task of the judge to indicate for each item the alternatives for which the borderline student should know that they are incorrect. The Nedelsky standard is then set equal to the sum of the reciprocals of the numbers of remaining alternatives of the items.

Let q_i denote the number of alternatives of item i and suppose that $k_i^{(c)}$ is the number of alternatives for which the judge indicates that a student with mastery level θ_c knows that they are incorrect. According to the model underlying the technique, the probability that this student answers item i correctly is equal to the reciprocal of the number of remaining alternatives:

$$P_i(+\,|\,\theta_c) = [q_i - k_i^{(c)}]^{-1}. \tag{4}$$

The Nedelsky standard is defined as:

$$\tau_c \equiv \sum_{i=1}^{n} [q_i - k_i^{(c)}]^{-1}. \tag{5}$$

Note that this is again a cutoff score on the true score scale since it is equal to the sum of probabilities used in Equation 3.

Obviously, the Angoff and Nedelsky techniques are based on different conceptions of the behavior that a student exhibits when responding to test items. In the Angoff technique, it is only assumed that his or her behavior is stochastic and that a student has different probabilities of success for different items. The Nedelsky technique supposes that a student proceeds by eliminating incorrect alternatives and then chooses at random between the remaining alternatives. The assumptions underlying the Angoff technique are extremely weak and consistent with the fact that behavioral measurements are subject to error. The Nedelsky technique, on the other hand, asserts that a specific behavior pattern can be expected, and this assertion can be true or false. That the Nedelsky technique is based on stronger assumptions is also clear from the range of values its probabilities of success can assume. In the Angoff technique, these probabilities can assume any value (between zero and one) but in the Nedelsky technique strong restrictions on the range of possible values are imposed and, in fact, only $q_i + 1$ different values are possible (namely the values $q^{-1}, (q-1)^{-1}, \ldots, 1$, and the value 0 representing the case in which the true alternative is eliminated). It is important to recognize that these values are unequally spaced and, in particular, that there are no probabilities possible between .5 and 1. Hence, it can be expected that there are many situations for which the Nedelsky technique does not hold but the Angoff technique still does.

LATENT TRAIT ANALYSIS

The probability of a successful item response, $P_i(+|\theta)$, varies as a function of the mastery level, θ. Generally, the higher the mastery level, the larger this probability. Latent trait theory is concerned with how $P_i(+|\theta)$ varies as a function of θ. It provides models for this function (usually called the item characteristic curve) and methods for analyzing their statistical properties and their fit to test data.

A versatile model in latent trait theory is the three-parameter logistic model:

$$P_i(+|\theta) = c_i + (1 - c_i)\{1 + \exp[-a_i(\theta - b_i)]\}^{-1}, \tag{6}$$

in which:

a_i is the discriminating power of item i,
b_i is the difficulty of item i,
c_i is a lower asymptote representing the probability of guessing item i correctly

(Birnbaum, 1968, pp. 399–405; Lord, 1980, pp. 12–14). The model is attractive because of its flexibility and the fact that it explicitly allows for such item properties as difficulty, discriminating power, and the possibility of guessing, all of which influence the probability of a successful response. For $a_i = 1$ and $c_i = 0$, the Rasch model (Rasch, 1960; Wright & Stone, 1979) is obtained. This model has unique, attractive statistical properties (Andersen, 1980, chap. 6) but is, due to the reduction of the number of parameters, less flexible than the model in Equation 6. More latent trait models are available. For further particulars, the reader should refer to the above references.

All latent trait models are approximations of the actual characteristic function of the items under consideration. If a model fits this function satisfactorily, it can be used for analyzing the item responses. For convenience, the model in Equation 6 will be used to describe how latent trait models can be employed for determining inconsistencies in the use of the Angoff and Nedelsky techniques. The choice is not essential, however; any other latent trait model could be used as well. The only important point is that each item

can be assumed to have a characteristic function showing how the probability of a successful response depends both on the mastery level of the student and the properties of the items.

METHOD

Suppose that Equation 6 holds for the n items for which the Angoff or Nedelsky technique is used. This implies that for each of these n items, a given level of mastery entails fixed probabilities of success. For example, suppose that the judge has a borderline student in mind who has a level of mastery equal to $\theta_c = .50$ and that one of the items has parameter values a = 1.25, b = .80, and c = .20. From Equation 6 it follows that the borderline student's probability of success on this item is equal to .53. However, the judge specifies a probability equal to .75. Obviously, an inconsistency has arisen since the judge's borderline student and the probability of success are incompatible with each other for the given item. Intrajudge inconsistency can also be illustrated using more than one item. Suppose that another item has a = .80, b = −.20, and c = .30 and that for this item the above judge specifies a probability of success equal to .30 for the borderline student. However, it follows from Equation 6 that these two probabilities of success imply different levels of mastery. The judge is inconsistent because probabilities of success are specified that can never belong to the *same* person.

As illustrated above, intrajudge inconsistencies arise when probabilities are specified that are incompatible with each other as well as with the borderline student the judge has in mind. This may be due to the fact that the properties of the items are not correctly perceived and, as a consequence, probabilities of success are specified that are not in accordance with the way the items actually function. The model given in Equation 6 suggests that the judge may have misperceived the difficulties or discriminating powers of the items or has inadequately allowed for the possibility of guessing.

For convenience, a somewhat different notation for the probabilities of success will be used. Let $p_i^{(s)}$ denote the borderline student's probability of success on item i as specified by an Angoff or Nedelsky judge. The superscript s is used to indicate that subjective probabilities are obtained. Further, the objective probabilities $P_i(+|\theta_c)$, which follow from the model given in Equation 6 for $\theta = \theta_c$, will be abbreviated as p_i. Now, a misspecification for item i occurs if:

$$e_i \equiv p_i^{(s)} - p_i \tag{7}$$

is nonzero. These quantities are of concern in the remaining part of the paper.

Note that the value of $p_i^{(s)}$ in Equation 7 is provided by the judge but that p_i must follow from the model. Hence, in order to use Equation 7, it is necessary to be able to estimate p_i. Estimates of p_i can be computed using Equation 6 provided that the item parameters have been estimated and θ_c is known. The former can be done with the aid of one of the available computer programs for parameter estimation. Concerning the latter, it is important to recall that the Angoff and Nedelsky techniques yield a cutoff score on the true score scale, τ, and that this scale is related to the θ scale via the test characteristic curve. Thus, the value of θ_c can be determined by computing τ_c from Equations 1 or 5 and using Equation 3 from the left to the right. Once this has been done, Equation 7 can be estimated for the test items to determine whether the judge has worked consistently enough to trust his/her probabilities of success. Note that the estimates of Equation 7 are obtained from the actual value of τ_c, i.e., under the hypothesis that the judge has worked

consistently, and that these estimates are subsequently used to decide whether this hypothesis is tenable. This is typical of hypothesis testing in statistics, where test statistics are derived under the assumption that the hypothesis to be tested holds.

In order to compare results between different judges or tests, an index of consistency may be useful. We could base such an index on the spread of θ values implied by the judge's probabilities. The larger this spread, the more inconsistent the probabilities. However, the θ-scale is only unique up to a linear transformation, making comparisons of measures of spread between different tests not always possible. A better choice, therefore, seems to choose an index of consistency based on the scale of the probabilities, i.e., on the standard interval [0, 1]. We first compute:

$$E \equiv \sum_{i=1}^{n} | p_i^{(s)} - p_i | / n, \tag{8}$$

which is the average absolute error of specification for the n items of the test. Note that the maximal size of the error $p_i^{(s)} - p_i$ in Equation 8 is restricted by the value of p_i and that the latter depends on θ_c. To be able to compare results between judges, however, the index has to be independent of the (arbitrary) borderline student's level of mastery the judge has in mind. Hence, a transformation is needed that makes Equation 8 free of its dependency on θ_c. This transformation should also reverse the scale of Equation 8 since in its present form that equation measures inconsistency instead of consistency.

A natural transformation is:

$$C_I \equiv \frac{M - E}{M}, \tag{9}$$

where:

$$M \equiv \sum_{i=1}^{n} e_i^{(u)} / n,$$

$$e_i^{(u)} \equiv \max \{p_i, 1 - p_i\}.$$

Note that $e_i^{(u)}$ is the maximum absolute value of the error of specification which follows when either $p_i^{(s)} = 0$ or $p_i^{(s)} = 1$ is substituted into Equation 7. C_I is thus the degree to which the average absolute error of specification deviates from its maximum possible value, measured on the standard interval (0, 1).

A special difficulty is associated with the use of the Nedelsky technique. As only $q_i + 1$ values for the probability of success on item i are possible, it follows from Equation 6 that no more than $q_i + 1$ values for θ_c are possible which each may differ from the mastery level the judge has in mind. Thus, for the Nedelsky technique, inconsistencies may be attributable not only to misperception by the judge but also to the discrete character of the technique. It is possible to estimate the loss of consistency due to the latter separately. As a result of the discrete character of the Nedelsky probabilities, the minimum value of the absolute error of specification, $| p_i^{(s)} - p_i |$, can be larger than zero. Generally, this minimum value is equal to:

$$e^{(\ell)} \equiv | [q_i - k_i^*]^{-1} - p_i |,$$

where k_i^* is the value of $k_i^{(c)}$ in Equation 4 chosen such that $e^{(\ell)}$ is minimal. Let:

$$m \equiv \sum_{i=1}^{n} e^{(\ell)} / n,$$

then:

$$C_2 = \frac{M - E}{M - m}$$

is a modification of C_1 allowing for the fact that for the Nedelsky technique, the smallest possible value of E is equal to m. Now,

$$\lambda = C_2 - C_1 \tag{10}$$

is equal to the reduction of consistency due to the discrete character of the technique. When using the Nedelsky technique, C_1 as well as λ should be estimated. C_1 indicates the consistency of the judge-Nedelsky technique combination. λ shows the reduction of consistency that can be ascribed to the Nedelsky technique; it can be used as a measure of the degree to which the model underlying the Nedelsky technique fits the situation.

The method proposed in this paper consists of two levels of model fitting. First, a latent trait model is fitted to item responses obtained from some set of examinees. Second, a standard or cutoff score is fitted on the latent variable using the probabilities of success provided by the judge. The latter is done under the hypothesis that the judge has worked consistently. Next, the errors of specification, $p_i^{(s)} - p_i$, are used to decide whether the hypothesis is in fact tenable.

More specifically, the method consists of the following steps:

1. A latent trait model is chosen, its parameters are estimated, and its fit is tested. Suppose that n items fit the model.
2. For these n items the Angoff or Nedelsky technique is used to specify for each item the probability of success $p_i^{(s)}$.
3. Using Equation 1 or 5, the Angoff or Nedelsky standard, τ_c, is computed.
4. The hypothesis to be tested is that the judge has worked consistently, i.e., has specified correct probabilities of success. Note that technically under this hypothesis, the Angoff or Nedelsky standard is a true score (expected observed score). The true score standard, τ_c, is next transformed into a standard on the θ-scale of the latent trait model via the estimated test characteristic curve, Equation 3. Since the latent trait standard θ_c is no explicit function of τ_c, trial values must be substituted for the former until the value of the latter is obtained. The task is simplified by the fact that θ is monotonically related to τ. However, some computer programs standardly produce the estimated test characteristic curve, and in that case $\hat{\theta}_c$ can simply be read off.
5. Next, substituting $\hat{\theta}_c$ and the estimated item parameters into the model, the estimated probabilities $\hat{p}_i$ are computed.
6. The index of consistency C_1 is computed to decide whether the judge has worked consistently. The closer the value of C_1 to zero, the less acceptable the hypothesis of a consistent judge.[1] If the conclusion is negative and the Nedelsky technique was used, λ can be computed to assess how large a reduction of consistency has occurred because of the discrete character of the technique.
7. Finally, the pattern of differences between $p_i^{(s)}$ and $\hat{p}_i$ is analyzed. Technically, these differences are the "residuals" left over after the hypothesis of a consistent judge has been fitted to the data. If the previous step has shown that the hypothesis is not

[1]The type I error for this decision cannot be specified as yet since no sensible model for the distribution of the $p_i^{(s)}$'s for a fixed judge and item seems possible.

tenable, then an analysis of this pattern is helpful in assessing where large misspecifications have occurred and where peculiarities in the judgments are present. The outcome of this analysis can be used, for instance, to detect items with systematic errors across judges or to determine for which type of items the judge needs additional training.

It is important that the results from step 6 be interpreted correctly. A value of C_1 close to one lends support to the assumption that the judge has specified consistent probabilities and that, therefore, the standard computed from these probabilities can be used safely. Low values of C_1 do not lend support to this assumption but, on the contrary, indicate that the judge has made errors in specifying the probabilities of success on the items. For short tests the occurrence of large errors implies an unreliable standard. It is not clear whether the same conclusion holds for longer tests and smaller errors. At first sight, one could argue that in this case the errors may have a tendency to average out. However, the situation differs from other estimation procedures in test theory, which are often unbiased by definition because the quantity to be estimated (e.g., the true score for a given person) is defined as the expected value of the observations. This does not hold here. It is simply unknown what standard would have been obtained when the judge had worked consistently, so it seems prudent not to trust standards for longer tests obtained from inconsistent judges.

RESULTS

An empirical investigation was carried out to illustrate the above method and to compare results for the Nedelsky and Angoff techniques. The items and Nedelsky data were taken from a previous investigation in which the values of item information functions at the Nedelsky standard were compared with pretest-posttest indices of item validity (van der Linden, 1981). All items were from a test belonging to the unit "Forces and Motion" from a physics course introducing tenth grade pupils to elementary mechanical concepts. The test was written by professional item writers of the National Institute of Educational Measurement, Arnhem, The Netherlands, in cooperation with the Project Team Curriculum Development Physics of the State University at Utrecht, The Netherlands. All items were of the three- and four-choice type. A latent trait analysis based on the responses of 156 pupils to an end-of-unit administration of the test produced 18 items showing a satisfactory fit to the Rasch model (Equation 6 with $a = 1$ and $c = 0$). For a further description of the test and the items, the reader is referred to van der Linden (1981).

The Nedelsky data were obtained using nine judges who all were involved in the curriculum development project. The judges were asked to conform to the learning objectives of the instructional unit as formulated in the project. The Angoff data were obtained for the same 18 items one year after the Nedelsky study took place, using eight different judges.

Table 1 shows the Nedelsky results for the nine judges. The first column gives the average absolute errors of specification. The next columns show the values for C_1, C_2 and λ. The mean error of specification in the whole study was .25. The mean value of λ in this study was equal to .09. As indicated earlier, this difference has to be explained by the lack of fit of the model underlying the Nedelsky technique.

In Table 2 the probabilities of success on all items are given both for the least consistent (Judge 2) and the most consistent judge (Judge 5). The first column contains the

Table 1

Results for Nine Judges Using the Nedelsky Technique

Judge	E	C_1	C_2	λ
1	.25	.65	.74	.09
2	.30	.63	.71	.08
3	.25	.65	.76	.11
4	.25	.69	.77	.08
5	.20	.75	.84	.09
6	.25	.69	.77	.08
7	.23	.69	.78	.09
8	.23	.73	.78	.05
9	.25	.67	.76	.09
Mean	.25	.68	.77	.09

Nedelsky probabilities which should be approximately equal to the estimated objective probabilities in the second column. For Judge 2 the differences between the two columns show large variability around their average absolute value of .30. These differences, albeit still considerable, are markedly smaller for Judge 5. The last two columns give the estimated values of $e^{(u)}$ and $e^{(\ell)}$ on which the computations of C_1 and λ are based. These values can also be used as benchmarks when inspecting the differences $p_i^{(s)} - p_i$ for the individual items.

The results for the Angoff technique are given in Table 3. As this table demonstrates, the average absolute errors are less serious than for the Nedelsky technique. The mean error in the whole study was equal to .18. Correspondingly, the values of C_1 are higher than those in Table 1.

Finally, Table 4 gives more detailed information for the most consistent and the least consistent Angoff judge. This information confirms that when using the Angoff or Nedelsky techniques one may have to reckon with serious misspecifications of the

Table 2

Estimated Probabilities of Success for Two Nedelsky Judges

	Judge 2				Judge 5			
Item	$p_i(s)$	$\hat{p}_i$	$\hat{e}_i(u)$	$\hat{e}_i(\ell)$	$p_i(s)$	$\hat{p}_i$	$\hat{e}_i(u)$	$\hat{e}_i(\ell)$
1	.50	.73	.73	.08	.33	.66	.66	.01
2	1.00	.11	.89	.12	.33	.08	.92	.08
3	1.00	.93	.93	.07	1.00	.90	.90	.10
4	.50	.50	.50	.16	.50	.41	.59	.04
5	1.00	.94	.94	.05	1.00	.92	.92	.08
6	.50	.84	.84	.15	.50	.79	.79	.12
7	1.00	.87	.87	.12	.50	.83	.83	.16
8	.50	.92	.92	.07	1.00	.89	.89	.11
9	.50	.71	.71	.05	.33	.63	.63	.04
10	.50	.86	.86	.13	.50	.81	.81	.14
11	.50	.74	.74	.01	.50	.67	.67	.08
12	.50	.16	.84	.08	.50	.12	.88	.12
13	.33	.82	.82	.17	1.00	.76	.76	.01
14	1.00	.22	.78	.01	.33	.17	.83	.08
15	.50	.26	.74	.02	.33	.20	.80	.05
16	.25	.62	.62	.12	.50	.53	.53	.03
17	1.00	.94	.94	.06	1.00	.91	.91	.09
18	.25	.17	.83	.07	.25	.13	.87	.12

Table 3

Results for Eight Judges Using the Angoff Technique

Judge	E	C_1
1	.21	.73
2	.15	.81
3	.16	.81
4	.20	.75
5	.16	.80
6	.17	.78
7	.22	.71
8	.19	.76
Mean	.18	.77

probabilities of success from which the standards are computed but that these are noticeably more favorable for the former than for the latter. It should be noted, however, that the results in Tables 1–4 have been obtained under specific conditions and that different results might have been found when, for instance, items of another difficulty or from other domains had been used.

DISCUSSION

Three possible sources of arbitrariness in standard setting using the Angoff or Nedelsky technique have been distinguished: (1) different conceptions of mastery underlying the technique, (2) different interpretations of the learning objectives, and (3) intrajudge inconsistency. It has been argued that differences in outcomes as a result of the first two sources can be expected and do not necessarily lead to unusable standards. What should be required is an explicit interpretation of the objectives as well as a conscious choice of the conception of mastery, both of which could be defended. The third source of arbitrariness can be serious, however. In this study errors of .20–.25 were typical but, especially for the Nedelsky technique, errors larger than .50 were no exception.

Table 4

Estimated Probabilities of Success for Two Angoff Judges

	Judge 2			Judge 7		
Item	$p_i(s)$	$\hat{p}_i$	$\hat{e}(u)$	$p_i(s)$	$\hat{p}_i$	$\hat{e}(u)$
1	.70	.74	.74	.30	.57	.57
2	.50	.11	.89	.30	.06	.94
3	.80	.93	.93	.90	.87	.87
4	.30	.50	.50	.70.	.34	.66
5	.80	.94	.94	.70	.89	.89
6	.90	.84	.84	.80	.72	.72
7	1.00	.87	.87	.50	.76	.76
8	.60	.92	.92	.30	.86	.86
9	.70	.72	.72	.30	.56	.56
10	.90	.86	.86	.60	.76	.76
11	.60	.75	.75	.70	.60	.60
12	.40	.16	.84	.30	.09	.91
13	.80	.82	.82	.50	.70	.70
14	.40	.23	.77	.50	.13	.87
15	.50	.27	.73	.30	.15	.85
16	.50	.62	.62	.50	.46	.54
17	.70	.94	.94	.80	.88	.88
18	.30	.18	.82	.50	.10	.90

The method proposed in this paper can be used for several purposes. An obvious possibility is a routine check of standard setting results before they are used in educational practice. Other possibilities are, for example: (1) selecting judges meeting predetermined criteria of consistency, (2) evaluating programs for training judges, (3) assessing the consequences of modifications of standard-setting techniques, or (4) item analysis to detect items yielding systematic errors across persons or techniques.

Another use of the method in this paper is for interactive standard setting. Work is in progress on an interactive computer program which is based on the idea that we should not let the judge work in the dark but assist him in reaching consistent probabilities for the items in the test. In this application of the method, the judge is asked to specify his success probabilities; the computer then confronts him with his inconsistencies and requests him to revise his probabilities. This is repeated until a consistent set of probabilities is reached.

For all these applications, it is necessary that items be available which fit one of the latent trait models. Two situations can arise. First, latent trait models can be used for item analysis in which items not fitting the model at first are revised or replaced until a test of appropriate length is obtained that fits the model. This procedure, albeit not always possible for practical reasons, is recommended because experience with latent trait analysis shows that items having unwanted properties are often not detected until such an analysis indicates that something is wrong. Moreover, all items are calibrated before the standard is set and the method proposed in the paper can immediately be used for the full test. A second situation arises if it is decided to use the method for a test and standard that are already in use. In this case, it may happen that some items do not fit the latent trait model satisfactorily, rendering the method usable only for the items that fit the model. In this event the procedure boils down to computing a new standard skipping the items not fitting the model and estimating Equations 7, 9, or 10 for the items that do fit the model. These estimates give an impression of how consistently the judge has worked.

REFERENCES

ANDERSEN, E. B. *Discrete statistical models with social science applications*. Amsterdam: North-Holland Publishing Company, 1980.

ANDREW, B. J., & HECHT, J. T. A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement*, 1976, **36**, 45–50.

ANGOFF, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1971.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

BRENNAN, R. L., & LOCKWOOD, R. E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, 1980, **4**, 219–240.

GLASS, G. V. Standards and criteria. *Journal of Educational Measurement*, 1978, **15**, 237–261.

HAMBLETON, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 1978, **15**, 277–290.

HAMBLETON, R. K. Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press, 1980.

HAMBLETON, R. K., POWELL, S., & EIGNOR, D. R. *Issues and methods for standard-setting.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, California, April 9–11, 1979.

HAMBLETON, R. K., SWAMINATHAN, H., ALGINA, R., & COULSON, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research,* 1978, **48,** 1–47.

JAEGER, R. M. Measurement consequences of selected standard-setting models. In M. A. Bunda and J. R. Sanders (Eds.), *Practices and problems in competency-based measurement.* Washington, D.C.: National Council on Measurement in Education, 1979.

KOFFLER, S. L. A comparison of approaches for setting standards. *Journal of Educational Measurement,* 1980, **17,** 167–178.

LORD, F. M. *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum, 1980.

NEDELSKY, L. Absolute grading standards for objective tests. *Educational and Psychological Measurement,* 1954, **14,** 3–19.

POPHAM, W. J. As always, provocative. *Journal of Educational Measurement,* 1978, **15,** 297–300.

RASCH, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmark Paedagogisk Institut, 1960.

SAUNDERS, J. C., RYAN, J. P., & HUYNH, H. A comparison of two approaches to setting passing scores based on the Nedelsky procedure. *Applied Psychological Measurement,* 1981, **5,** 209–217.

SHEPARD, L. A. Setting standards. In M. A. Bunda and J. R. Sanders (Eds.), *Practices and problems in competency-based measurement.* Washington, D.C.: National Council on Measurement in Education, 1979.

SHEPARD, L. A. Standard setting issues and methods. *Applied Psychological Measurement,* 1980, **4,** 447–467. (a)

SHEPARD, L. A. Technical issues in minimum competency testing. In D. C. Berliner (Ed.), *Review of research in education* (Vol. 8). Itasca, IL: F. E. Peacock Publishers, 1980. (b)

SKAKUN, E. N., & KLING, S. Comparability of methods for setting standards. *Journal of Educational Measurement,* 1980, **17,** 229–235.

VAN DER LINDEN, W. J. Decision models for use with criterion-referenced tests. *Applied Psychological Measurement,* 1980, **4,** 469–492.

VAN DER LINDEN, W. J. A latent trait look at pretest-posttest validation of criterion-referenced test items. *Review of Educational Research,* 1981, **51,** 379–402.

VAN DER LINDEN, W. J. Criterion-referenced measurement: Its main applications problems, and findings. In W. J. van der Linden (Ed.), Aspects of criterion-referenced measurement. *Evaluation in Education: An International Review Series,* 1982, **6,** 97–118.

WRIGHT, B. D., & STONE, M. H. *Best test design: A handbook for Rasch measurement.* Chicago, IL: MESA Press, 1979.

AUTHOR

VAN DER LINDEN, WIM J. *Address:* Department of Education, Twente University of Technology, 7500 AE Enschede, The Netherlands. *Title:* Senior Lecturer. *Degrees:* B. Psych., B. Soc., M. Psych., M. Soc., University of Utrecht, Ph. D. University of Amsterdam. *Specialization:* Psychometric methods, Data analysis, Research methodology.